Spatial and Angular Resolution Enhancement of Light Fields Using Convolutional Neural Networks

M. Shahzeb Khan Gul and Bahadir K. Gunturk
Dept. of Electrical and Electronics Engineering, Istanbul Medipol University, Istanbul, Turkey
mskhangul@st.medipol.edu.tr, bkgunturk@medipol.edu.tr

***Abstract*—Light field imaging extends the traditional photography by capturing both spatial and angular distribution of light, which enables new capabilities, including post-capture refocusing, post-capture aperture control, and depth estimation from a single shot. Micro-lens array (MLA) based light field cameras offer a cost-effective approach to capture light field. A major drawback of MLA based light field cameras is low spatial resolution, which is due to the fact that a single image sensor is shared to capture both spatial and angular information. In this paper, we present a learning based light field enhancement approach. Both spatial and angular resolution of captured light field is enhanced using convolutional neural networks. The proposed method is tested with real light field data captured with a Lytro light field camera, clearly demonstrating spatial and angular resolution improvement.**

***Index Terms*—Light field, super-resolution, convolutional neural network.**

I. Introduction

Light field refers to the collection of light rays in 3D space. With a light field imaging system, light rays in different directions are recorded separately, unlike a traditional imaging system, where a pixel records the total amount of light received by the lens regardless of the direction. The angular information enables new capabilities, including depth estimation, post-capture refocusing, post-capture aperture size and shape control, and 3D modelling. Light field imaging can be used in different application areas, including 3D optical inspection, robotics, microscopy, photography, and computer graphics.

Light field imaging is first described by Lippmann, who proposed to use a set of small biconvex lenses to capture light rays in different directions and refers to it as integral imaging [1]. The term "light field" was first used by Gershun, who studied the radiometric properties of light in space [2]. Adelson and Bergen used the term "plenoptic function" and defined it as the function of light rays in terms of intensity, position in space, travel direction, wavelength, and time [3]. Adelson and Wang described and implemented a light field camera that incorporates a single main lens along with a micro-lens array [4]. This design approach is later adopted in commercial light field cameras [5], [6]. In 1996, Levoy and Hanrahan [7] and Gortler *et al.* [8] formulated light field as a 4D function, and studied ray space representation and light field re-sampling. Over the years, light field imaging theory and applications have continued to be developed further. Key developments include post-capture refocusing [9], Fourier-domain light field processing [10], light field microscopy [11], focused plenoptic camera [12], and multi-focus plenoptic camera [13].

Light field acquisition can be done in various ways, such as camera arrays [14], optical masks [15], angle-sensitive pixels [16], and micro-lens arrays [10], [12]. Among these different approaches, micro-lens array (MLA) based light field cameras provide a cost-effective solution, and have been successfully commercialized [5], [6]. There are two basic implementation approaches of MLA-based light field cameras. In one approach, the image sensor is placed at the focal length of the micro-lenses [10], [5]. In the other approach, a micro-lens relays the image (formed by the objective lens on an intermediate image plane) to the image sensor [12], [6]. These two approaches are illustrated in Figure 1. In the first approach, the sensor pixels behind a micro-lens (also called a lenslet) on the MLA record light rays coming from different directions. Each lenslet region provides a single pixel value for a perspective image; therefore, the number of lenslets corresponds to the number of pixels in a perspective image. That is, the spatial resolution is defined by the number of lenslets in the MLA. The number of pixels behind a lenslet, on the other hand, defines the angular resolution, that is, the number of perspective images. In the second approach, a lenslet forms an image of the scene from a particular viewpoint. The number of lenslets defines the angular resolution; and, the number of pixels behind a lenslet gives the spatial resolution of a perspective image.

In the MLA-based light field cameras, there is a trade-off between spatial resolution and angular resolution, since a single image sensor is used to capture both. For example, in the first generation Lytro camera, an 11 megapixel image sensor produces 11x11 sub-aperture perspective images, each with a spatial resolution of about 0.1 megapixels. Such a low spatial resolution prevents the widespread adoption of light field cameras. In recent years, different methods have been proposed to tackle the low spatial resolution issue. Hybrid systems, consisting of a light field sensor and a regular sensor, have been presented [17], [18], [19], where the high spatial resolution image from the regular sensor is used to enhance the light field sub-aperture (perspective) images. The disadvantages of hybrid systems include increased cost and larger camera dimensions. Another approach is to apply multi-frame super-resolution techniques to the sub-aperture images of a light field [20], [21]. It is also possible to apply learning-based super-resolution techniques to each sub-aperture image of a light field [22].

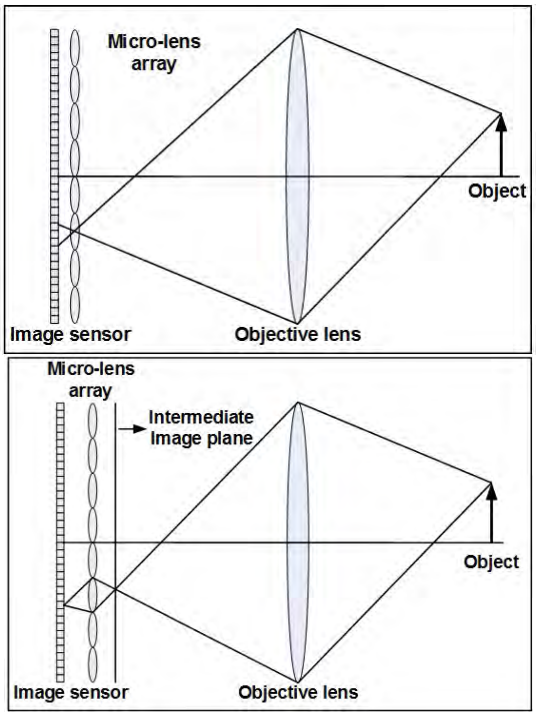


Fig. 1: Two main approaches for MLA-based light field camera design. Top: The distance between the image sensor and the MLA is equal to the focal length of a micro-lens (lenslet) in the MLA. Bottom: The objective lens forms an image of the scene on an intermediate image plane, which is then relayed by the lenslets to the image sensor.

In this paper, we present a convolutional neural network based light field super-resolution method. The method has two sub-networks; one is trained to increase the angular resolution, that is, to synthesize novel viewpoints (sub-aperture images); and the other is trained to increase the spatial resolution of each sub-aperture image. We show that the proposed method provides significant increase in image quality, visually as well as quantitatively (in terms of peak signal-to-noise ratio and structural similarity index [23]), and improves depth estimation accuracy.

The paper is organized as follows. We present the related work in the literature in Section II. We explain the proposed method in Section III, present our experimental results in Section IV, and conclude the paper in Section V.

II. Related Work

A. Super-Resolution of Light Field

One approach to enhance the spatial resolution of images captured with an MLA-based light field camera is to apply a multi-frame super-resolution technique on the perspective images obtained from the light field capture. The Bayesian super-resolution restoration framework is commonly used, with Lambertian and textual priors [20], Gaussian mixture models [24], and variational models [21].

Learning-based single-image super-resolution methods can also be adopted to address the low spatial resolution issue of light fields. In [22], a dictionary learning based super-resolution method is presented, demonstrating a clear improvement over standard interpolation techniques when converting raw light field capture into perspective images. Another learning based method is presented in [25], which incorporates deep convolutional neural networks for spatial and angular resolution enhancement of light fields. Alternative to spatial domain resolution enhancement approaches, frequency domain methods, utilizing signal sparsity and Fourier slice theorem, have also been proposed [26], [27].

In contrast to single-sensor light field imaging systems, hybrid light field imaging system have also been introduced to improve spatial resolution. In the hybrid imaging system proposed by Boominathan *et al.* [17], a patch-based algorithm is used to super-resolve low-resolution light field views using high-resolution patches acquired from a standard high-resolution camera. There are several other hybrid imaging system presented [18], [28], [19], combining images from a standard camera and a light field camera. Among these, the work in [19] demonstrates a wide baseline hybrid stereo system, improving range and accuracy of depth estimation in addition to spatial resolution enhancement.

B. Deep Learning for Image Restoration

Convolutional neural networks (CNNs) are variants of multi-layer perceptron networks. Convolution layer, which is inspired from the work of Hubel and Wiesel [29] showing that visual neurons respond to local regions, is the fundamental part of a CNN. In [30], LeCun *et al.* presented a convolutional neural network based pattern recognition algorithm, promoting further research in this field. Deep learning with convolutional neural networks has been extensively and successfully applied to computer vision applications. While most of these applications are on classification and object recognition, there are also deep-learning based low-level vision applications, including compression artifact reduction [31], image deblurring [32] [33], image deconvolution [34], image denoising [35], image inpainting [36], removing dirt/rain noise [37], edge-aware filters [38], image colorization [39], and in image segmentation [40]. Recently, CNNs are also used for super-resolution enhancement of images [41], [42], [43], [44]. Although these single-frame super-resolution methods can be directly applied to light field perspective images to improve their spatial resolution, we expect better performance if the angular information available in the light field data is also exploited.

III. Light Field Super Resolution using Convolutional Neural Network

In Figure 2, a light field captured by a micro-lens array based light field camera (Lytro Illum) is shown. When zoomed-in, individual lenslet regions of the MLA can be seen. The pixels behind a lenslet region record directional light intensities received by that lenslet. As illustrated in Figure 3, it is possible to represent a light field with four parameters *(s,t,u,v)*, where *(s,t)* indicates the lenslet location, and *(u,v)* indicates the angular position behind the lenslet. A perspective

Fig. 2: Light field captured by a Lytro Illum camera. A zoomed-in region is overlaid to show the individual lenslet regions.

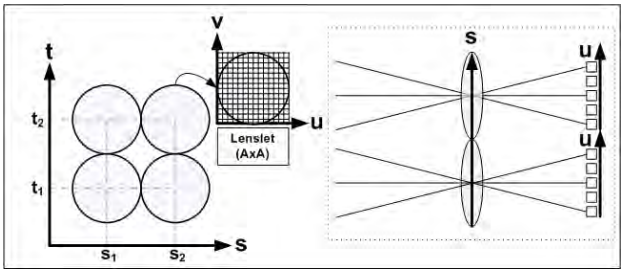


Fig. 3: Light field parameterization. Light field can be parameterized by the lenslet positions *(s,t)* and the pixel positions *(u,v)* behind a lenslet.

image can be constructed by taking a single pixel value with a specific *(u,v)* index from each lenslet. The process is illustrated in Figure 4. The spatial resolution of a perspective image is controlled by the size and the number of the lenslets. Given a fixed image sensor size, the spatial resolution can be increased by having smaller size lenslets; given a fixed lenslet size, the spatial resolution can be increased by increasing the number of lenslets, thus, the size of the image sensor. The angular resolution, on the other hand, is defined by the number of pixels behind a lenslet region.

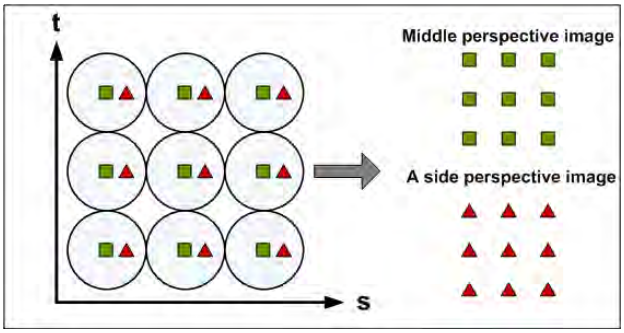


Fig. 4: Sub-aperture (perspective) image formation. A perspective image can be constructed by picking specific pixels from the lenslet regions. The size of a perspective image is determined by the number of lenslets.

Our goal is to increase both spatial and angular resolution of a light field capture. We propose a convolutional neural network based learning method, which we call *light field super resolution* (LFSR). It consists of two steps. Given a light field where there are AxA pixels in each lenslet area and the size of each perspective is HxW, the first step doubles the angular resolution from AxA to $2A$x$2A$ using a convolutional neural network. In the second step, the spatial resolution is doubled from HxW to $2H$x$2W$ by estimating new lenslet regions between given lenslet regions. Figure 5 gives an illustration of these steps.

The closest work in the literature to our method is the one presented in [25], which also uses deep convolutional networks. There is a fundamental difference between our approach and the one in [25]; while our architecture is designed to work on raw light field data, that is, lenslet regions; [25] is designed to work on perspective images. In the experimental results section, we provide both visual and quantitative comparisons with [25].

A. Angular Super-Resolution (SR) Network

The proposed angular super-resolution network is shown in Figure 6. It is composed of two convolution layers and a fully connected layer. The input to the network is a lenslet region with size AxA; and the output is a higher resolution lenslet region with size $2A$x$2A$. That is, the angular resolution enhancement is done directly on the raw light field (after demosaicking) as opposed to doing on perspective images. Each lenslet region is interpolated by applying the same network. Once the lenslet regions are interpolated, one can construct the perspective images by rearranging the pixels, as mentioned before. At the end, $2A$x$2A$ perspective images are obtained from AxA perspective images.

The convolution layers in the proposed architecture are based on the intuition that the first layer extracts a high-dimensional feature vector from the lenslet and the second convolution layer maps it onto another high-dimensional vector. After each convolution layer, there is a non-linear activation layer of *Rectified Linear Unit* (ReLU). In the end, a fully connected layer aggregates the information of the last convolution layer and predicts a higher-resolution version of the lenslet region.

The first convolution layer has n_1 filters, each with size n_0xk_1xk_1. (In our experiments, we treat each color channel separately, thus $n_0 = 1$.) The second convolution layer has n_2 filters, each with size n_1xk_2xk_2. The final layer is a fully connected layer with $4A^2$ neurons, forming a $2A$x$2A$ lenslet region.

B. Spatial Super-Resolution (SR) Network

Figure 7 gives an illustration of the spatial super-resolution network. Similar to the angular super-resolution network, the architecture has two convolution layers, each followed by a ReLU layer, followed by a fully connected layer. Different from the angular resolution network, four lenslet regions are stacked together as the input to the network. There are three

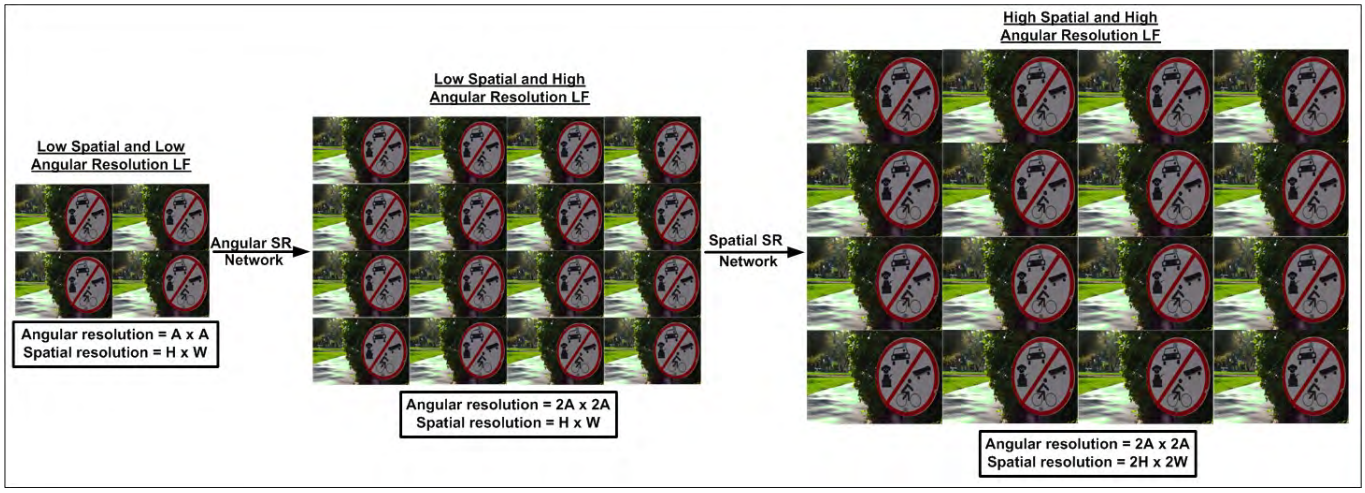


Fig. 5: An illustration of the proposed LFSR method. First, the angular resolution of the light field (LF) is doubled; second, the spatial resolution is doubled. The networks are applied directly on the raw demosaicked light field, not on the perspective images.

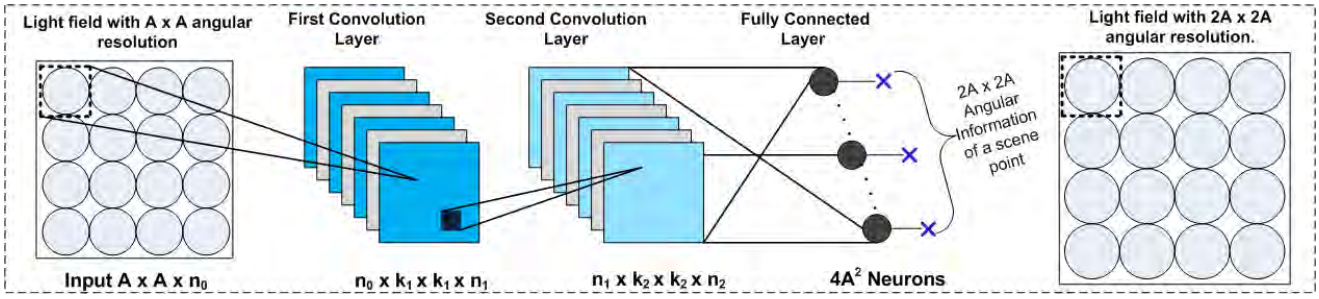


Fig. 6: Overview of the angular SR network to estimate a higher-angular resolution version of the input light field. A lenslet is drawn as a circle; the AxA region behind a lenslet is taken as the input and processed to predict the corresponding $2A$x$2A$ lenslet region. Each convolution layer is followed by a non-linear activation layer of ReLU.

outputs at the end, predicting the horizontal, vertical, and diagonal sub-pixels of a perspective image. To clarify the idea further, Figure 8 illustrates the formation of a high-resolution perspective image. As mentioned earlier, a perspective image of a light field is formed by picking a specific pixel from each lenslet region and putting all picked pixels together according to their respective lenslet positions. Using four lenslet regions, the network predicts three additional pixels in between the pixels picked from the lenslet regions. The predicted pixels, along with the picked pixels, form a higher resolution perspective image.

C. Training the Networks

We used a dataset that is captured by a Lytro Illum camera [45]. The dataset has more than 200 raw light fields, each with an angular resolution of 14x14 and a spatial resolution of 374x540. In other words, each light field consists of 14x14 perspective images; and each perspective image has a spatial resolution of 374x540 pixels. The raw light field is of size 5236x7560, consisting of 374x540 lenslet regions, where each lenslet region has 14x14 pixels. We used 45 light fields for training and reserved the others for testing. The training data is obtained in two steps. First, we drop every other lenslet region to obtain a low-spatial-resolution (187x270) and high-angular-resolution (14x14) light field. Second, we drop every other pixel in a lenslet region to obtain a low-spatial-resolution (187x270) and low-angular-resolution (7x7) light field.

The angular SR network, as shown in Figure 6, has low-spatial-resolution and low-angular-resolution light field as its input, and low-spatial-resolution and high-angular-resolution light field as its output. Each lenslet region is treated separately by the network, increasing the size from 7x7 to 14x14. The first convolution layer consists of 64 filters, each with size 1x3x3. It is followed by a ReLU layer. The second convolution layer consists of 32 filters of size 64x1x1, followed by a ReLU layer. Finally, there is a fully connected layer with 196 neurons to produce a 14x14 lenslet region.

The spatial SR network, as shown in Figure 7, has low-spatial-resolution and high-angular-resolution light field as its input, and high-spatial-resolution and high-angular-resolution light field as its output. Four lenslet regions are stacked to form a 14x14x4 input. The first convolution layer consists of 64 filters, each with size 4x3x3. The second convolution layer consists of 32 filters of size 64x1x1. Each convolution layer is followed by a ReLU layer. Finally, there is a fully connected layer with three neurons to produce the horizontal, vertical

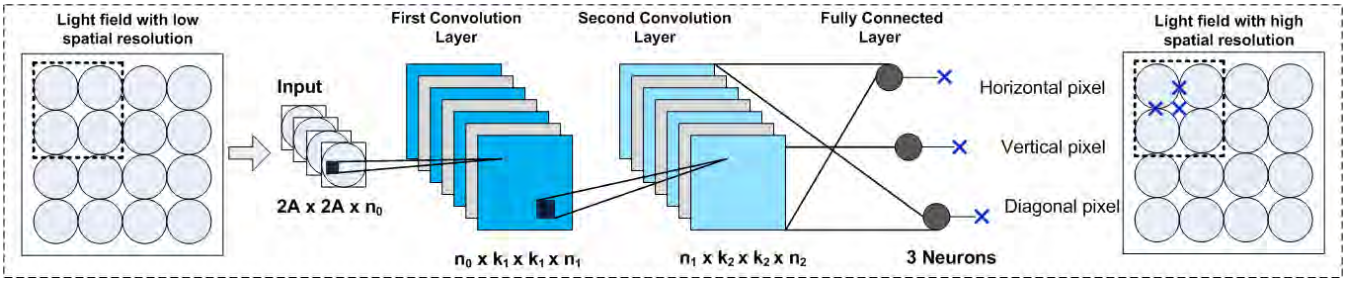


Fig. 7: Overview of the proposed spatial SR network to estimate a higher-spatial resolution version of the input light field. Four lenslet regions are stacked and given as the input to the network. The network predicts three new pixels to be used in the high-resolution perspective image formation. Each convolution layer is followed by a non-linear activation layer of ReLU.

TABLE I: Comparison of different spatial and angular resolution enhancement methods.

Methods	PSNR (dB)			SSIM		
	Min	Avg	Max	Min	Avg	Max
Bicubic resizing (*imresize*)	24.2029	27.6671	34.6330	0.7869	0.8744	0.9457
LFCNN [25]	25.5963	28.9661	34.8231	0.7838	0.8904	0.9407
Bicubic interpolation	27.2620	30.6245	37.1640	0.5780	0.9256	0.9659
Proposed (LFSR)	**29.7515**	**33.4273**	**39.5655**	**0.9360**	**0.9559**	**0.9823**

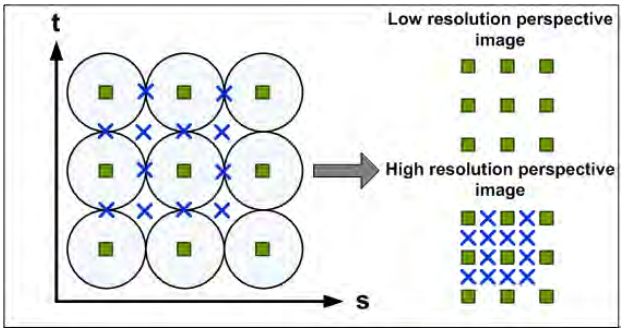


Fig. 8: Constructing a high-resolution perspective image. A perspective image can be formed by picking a specific pixel from each lenslet region, and putting all picked pixels together. Using the additional pixels predicted by the spatial SR network, a higher-resolution perspective image is formed.

and diagonal pixels. This network generates one high-spatial resolution perspective. For each perspective, the network is trained separately.

We implement and train our model using the Caffe package [46]. For the weight initialization of both networks, we used the initialization technique given in [47], with mean value set to zero and standard deviation set to 10^{-3}, to prevent vanishment or over-amplification of weights. The learning rates for the three layers of the networks are 10^{-3}, 10^{-3}, and 10^{-5}, respectively. Mean squared error is used as the loss function, which is minimized using the stochastic gradient descent method with standard backpropagation [30]. For each network, the input size is about 13 million; and the number of iterations is about 10^8.

IV. Experiments

We evaluated our LFSR method on 25 test light fields which we reserved from the Lytro Illum camera dataset [45] and on the HCI dataset [48]. For spatial and angular resolution enhancement, we compared our method against the LFCNN [25] method and bicubic interpolation. There are several methods in the literature that synthesize new viewpoints from a light field data; thus, we compared the angular SR network of our method with two such view synthesis methods, namely, Kalantari *et al.* [49] and Wanner and Goldluecke [50]. Finally, there are single-frame spatial resolution enhancement methods; we chose the latest state-of-the-art method, called DRRN [51], and included it in our comparisons.

In addition to spatial and angular resolution enhancement, we investigated depth estimation performance, and compared the depth maps generated by low-resolution light fields and the resolution-enhanced light fields. In the end, we investigated the effect of the network parameters, including the filter size and the number of layers, on the performance of the proposed spatial SR network.

A. Spatial and Angular Resolution Enhancement

The test images are downsampled from 14x14 perspective images, each with size 374x540 pixels, to 7x7 perspective images with size 187x270 pixels by dropping every other lenslet region and every pixel in each lenslet region. The trained networks are applied to these low-spatial and low-angular resolution images to bring them back to the original spatial and angular resolutions. The networks are applied on each color channel separately. Since the original perspective images available, we can quantitatively calculate the performance by comparing the estimated and the original images. In Table I, we provide peak-signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [23] results of our method, in addition to the results of the LFCNN [25] method and bicubic interpolation. Here, we should make two notes about the LFCNN method. First, we took the learned parameters provided in the original paper and fine tuned them with our dataset as described in [25]. This revision improves the

performance of the LFCNN method for our dataset. Second, the LFCNN method is designed to split a low-resolution image pixel into four sub-pixels to produce a high-resolution image; therefore, we included the results of bicubic resizing (*imresize* function in MATLAB) to evaluate the quantitative performance of the LFCNN method. In Table I, we see that the LFCNN method produces about 1.3 dB better than the bicubic resizing. The proposed method produces the best results in terms of PSNR and SSIM.

Visual comparison is critical when evaluating spatial resolution enhancement. Figures 9, 10, and 11 are typical results from the test dataset. Figure 12 is our worst result among all test images. In these figures, we also include the results of the single-image spatial resolution method, called DRRN [51]. This method is based on deep recursive residual network technique, and produces state-of-the-art results in spatial resolution enhancement. Examining the results visually, we conclude that our method performs better the LFCNN method and bicubic interpolation, and produces comparable results with the DRNN method. We notice that the LFCNN method produces sharper results compared to bicubic interpolation despite having lower PSNR values. In our worst result, given in Figure 12, the DRNN method outperforms all methods. This particular image has highly complex texture, which seems to be not modelled well with the proposed architecture. Training with similar images or using more complex architecture may improve the performance. When comparing deep networks, we should consider the computational cost as well. The computation time for one image with the DRRN method is about 859 seconds, whereas, the proposed SR network takes about 53 seconds, noting that both are implemented in MATLAB on the same machine.

In Figure 13, we test our method on the HCI dataset [48]. We compare against the networks in [25] and [52]. The method in [52] produces less ringing artifacts compared to the LFCNN network [25]. The proposed method again produces the best visual results.

Although we have showed results for resolution enhancement of the middle perspective image so far, the proposed spatial SR network can be used for any perspective image as well. In Table II, average PSNR and SSIM on test images for different perspective images (among the 14x14 set) are presented. It is seen that similar results are obtained on all perspective images, as expected.

TABLE II: Evaluation of the proposed method for different perspective images.

Perspective image (Row #, Column #)	Method	PSNR (dB)	SSIM
(7,1)	Bicubic interp.	28.21	0.8631
	Proposed	**28.74**	**0.8789**
(5,5)	Bicubic interp.	31.12	0.9310
	Proposed	**32.95**	**0.9496**
(6,6)	Bicubic interp.	30.74	0.9272
	Proposed	**32.71**	**0.9485**
(6,8)	Bicubic interp.	30.73	0.9267
	Proposed	**32.94**	**0.9504**
(8,6)	Bicubic interp.	30.69	0.9270
	Proposed	**32.69**	**0.9492**
(8,8)	Bicubic interp.	27.71	0.8793
	Proposed	**28.16**	**0.8917**

B. Angular Resolution Enhancement

In this section, we evaluate the individual performance of our angular SR network. For this experiment, the angular resolution of the test images are downsampled from 14x14 to 7x7 while keeping the spatial resolution at 374x540 pixels. These low-angular images are then input to the angular SR network to bring them back to the original angular resolution. The network is trained for each color channel separately. We compare our method against Kalantari *et al.* [49], which is a very recent convolutional neural network based novel view synthesis method, and against Wanner and Goldluecke [50], which utilizes disparity maps in a variational optimization framework. Wanner and Goldluecke [50] may work with any disparity map generation algorithm; thus, we report results with the disparity generation algorithms given in [53], [54], [55], and [56]. In Table III, we quantitatively compared the results with the state-of-the-art angular resolution enhancement methods using PSNR and SSIM. In Figure 14, we provide a visual comparison. The scene contains occluded regions, which are generally difficult for view synthesis. Our angular SR method produces significantly better results compared to all other approaches.

Finally, we would like to note that the angular SR network, by itself, may turn out to be useful, since it may be combined with any single-image resolution enhancement method to enhance the spatial and angular resolution of a light field capture.

C. Depth Map Estimation Accuracy

One of the capabilities of light field imaging is depth map estimation, whose accuracy is directly related to the angular resolution of light field. In Figure 15 and Figure 16, we compare depth maps obtained from the input light fields and the light fields enhanced by the proposed method. The depth maps are estimated using the method in [56], which is specifically designed for light fields. It is clearly seen that depth maps obtained from light fields enhanced with the proposed method show higher accuracy. With the enhanced light fields, even close depths can be differentiated, unlike the low-resolution light fields.

D. Model and Performance Trade-Offs

To evaluate the trade-off between performance and speed, and to investigate the relation between performance and the network parameters, we modify different parameters of the network architecture and compare with the base architecture. All the experiments are performed on a machine with Intel Xeon CPU E5-1650 v3 3.5GHz, 16GB RAM and Nvidia 980ti 6GB graphics card.

Filter Size: In the proposed spatial SR network, the filter sizes in the two convolution layers are $k_1 = 3$ and $k_2 = 1$, respectively. The filter size of the first convolution layer is kept at $k_1 = 3$; this means, for each light ray (equivalently, perspective image), the network is considering the light rays (perspective

(a) Ground truth (b) Bicubic resizing (*imresize*) / 25.34 dB (c) Bicubic interp. / 27.7 dB (d) LFCNN [25] / 25.76 dB (e) DRRN [51] / 31.63 dB (f) LFSR / 32.25 dB

Fig. 9: Visual comparison of different methods.

TABLE III: Comparison of different methods for angular resolution enhancement.

Picture name	Evaluation metric	Wanner and Goldluecke [50] Disparity [53]	Wanner and Goldluecke [50] Disparity [54]	Wanner and Goldluecke [50] Disparity [55]	Wanner and Goldluecke [50] Disparity [56]	Kalantari *et al.* [49]	Angular SR network
Flower 1	PSNR (dB)	22.03	29.52	24.39	28.21	33.31	**35.95**
	SSIM	0.789	0.941	0.910	0.934	0.969	**0.982**
Cars	PSNR (dB)	19.74	27.27	22.09	27.51	31.65	**35.21**
	SSIM	0.792	0.946	0.911	0.949	0.966	**0.983**
Flower 2	PSNR (dB)	20.61	27.56	23.65	27.04	31.93	**36.75**
	SSIM	0.645	0.919	0.899	0.924	0.959	**0.980**
Rock	PSNR (dB)	16.57	30.46	30.55	30.21	**34.67**	34.09
	SSIM	0.488	0.945	0.948	0.946	**0.970**	0.963
Leaves	PSNR (dB)	15.03	23.54	20.08	23.88	27.80	**33.08**
	SSIM	0.481	0.882	0.855	0.893	**0.963**	0.956

images) in a 3x3 neighborhood in the first convolution layer. Since higher dimensional relations are taken care of in the second convolution layer, and since keeping the filter size small minimizes the boundary effects—note that the input size in the first layer is 14x14—this seems to be a reasonable choice for the first layer. On the other hand, we have more flexibility in the second convolution layer. We examined the effect of the filter size in the second convolution layer by setting k_2 = 3 and k_2 = 5 while keeping the other parameters intact. In Figure 17, we provide the average PSNR values on the test dataset for different values of k_2 as a function of training backpropagation numbers. When $k_2 = 5$, the convergence is slightly better than the case with $k_2 = 1$. In Table IV, we show the final PSNR values and the computation times per channel (namely, the red channel) for a perspective image. It is seen that while the PSNR is slightly improved the computation time is more than doubled when we increase the filter size from $k_2 = 1$ to $k_2 = 5$.

TABLE IV: Effect of the filter size on performance and the speed of the spatial SR network.

Filter size	PSNR (dB)	Time (sec)
k_2 = 1	36.31	17.58
k_2 = 3	35.81	27.62
k_2 = 5	36.40	45.18

(a) Ground truth (b) Bicubic resizing (*imresize*) / 25 dB (c) Bicubic interp. / 28.11 dB (d) LFCNN [25] / 25.12 dB (e) DRRN [51] / 32.20 dB (f) LFSR / 32.35 dB

Fig. 10: Visual comparison of different methods.

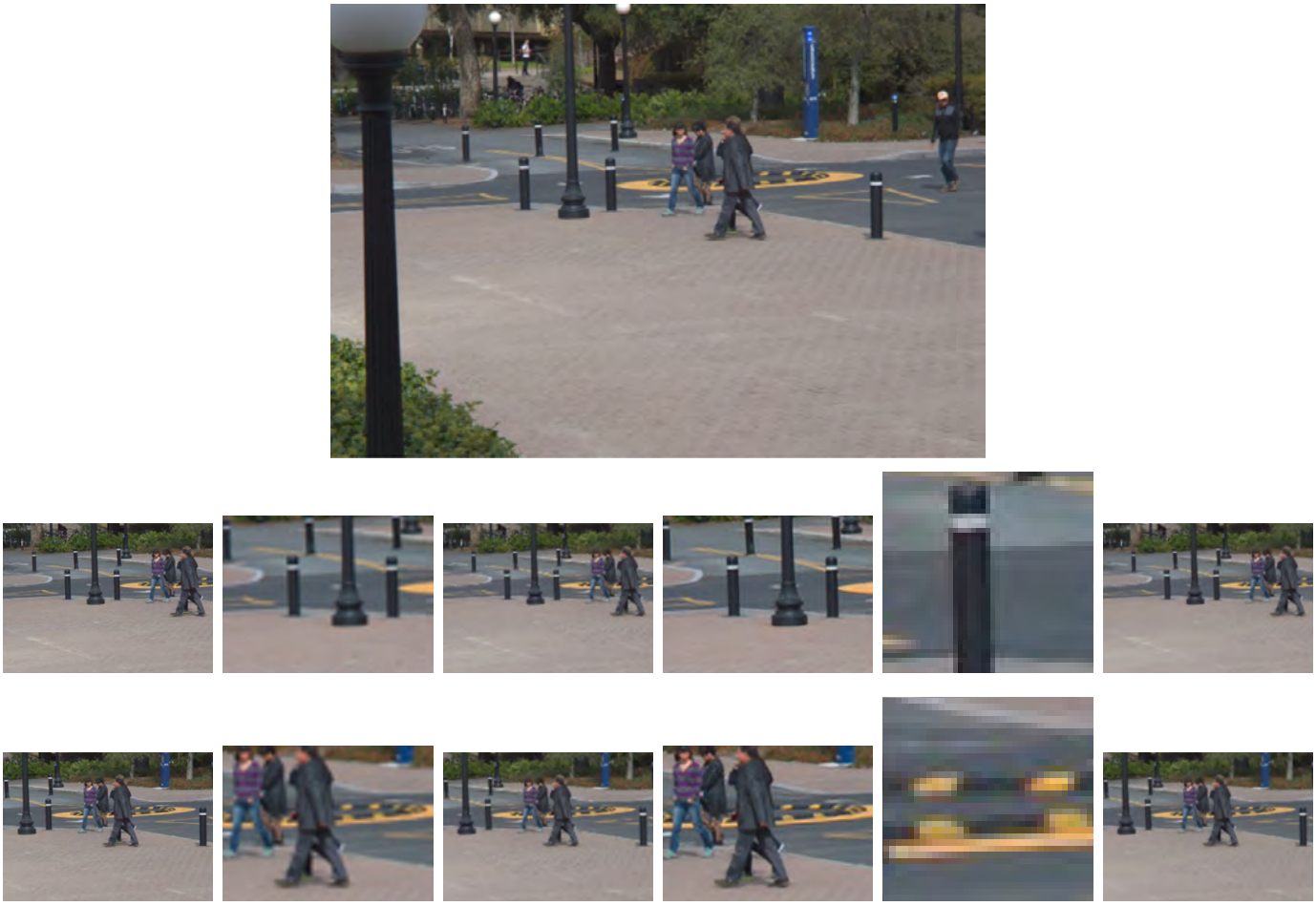
Number of Layers and Number of Filters: We also examine the network performance for different number of layers and different number of filters. We implemented deeper architectures by adding new convolution layers after the second convolution layer. The three-layer network presented in the previous section is compared against the four-layer and five-layer networks. For the four-layer network, we evaluated the performance for different filter combinations. The network configurations we used are shown in Table V. In Figure 18, we provide the convergence curves for these different network configurations. We observe that the simple three-layer network performs better than the others. This means that increasing the number of convolution layers is causing overfitting and degrading the performance.

TABLE V: Different network configurations used to evaluate the performance of the spatial SR network.

	Convolution Layers			
	First	Second	Third	Fourth
3 layer	1x3x3x64	64x1x1x32	-	-
4 layer	1x3x3x64	64x1x1x32	32x1x1x32	-
4 layer	1x3x3x64	64x1x1x16	16x1x1x16	-
4 layer	1x3x3x64	64x1x1x32	32x1x1x16	-
5 layer	1x3x3x64	64x1x1x16	16x1x1x16	16x1x1x16

E. Further Increasing the Spatial Resolution

For quantitative evaluation, we need to have the ground truth; thus, we downsample the captured light field to generate its lower resolution version. In addition, we can visually evaluate the performance of the proposed method without downsampling and further increasing the spatial resolution of the original images. In Figure 19, we provide a comparison of bicubic resizing, bicubic interpolation, the LFCNN method [25], the DRRN method [51], and the proposed LFSR method. The spatial resolution of each perspective image is increased from 374x540 to 748x1080. The results of the proposed method seem to be preferable over the others with less artifacts. The LFCNN results in sharp images but has some visible artifacts. The DRNN method seems to distort some texture, especially visible in the second example image, while the proposed method preserves the texture well.



(a) Ground truth (b) Bicubic resizing (*imresize*) / 28.92 dB (c) Bicubic interp. / 28.11 dB (d) LFCNN [25] / 29.06 dB (e) DRRN [51] / 36.78 dB (f) LFSR / 34.21 dB

Fig. 11: Visual comparison of different methods.

V. Discussion and Conclusions

In this paper, we presented a convolutional neural network based light field super-resolution method. The method consists of two separate convolutional neural networks trained through supervised learning. The architecture of these networks are composed of only three layers, reducing computational complexity. The proposed method shows significant improvement both quantitatively and visually over the baseline bicubic interpolation and another deep learning based light field super-resolution method. In addition, we compared the angular resolution enhancement part of our method against two methods for novel view synthesis. We also demonstrated that enhanced light field results in more accurate depth map estimation due to the increase in angular resolution.

The spatial super-resolution network is designed to generate one perspective image. One may suggest to generate all perspectives in a single run; however, this would result in a larger network, requiring larger size dataset and more training. Instead, we preferred to have a simple, specialized, and effective architecture.

Similar to other neural network based super-resolution techniques, the method is designed to increase the resolution by an integer factor (two). It can be applied multiple times to increase the resolution by factors of two. A non-integer factor size change is also possible by first interpolating using the proposed method and then downsampling using a standard technique.

The network parameters are optimized for a specific light field camera. For different cameras, the specific network parameters, such as filter dimensions, may need to be optimized. We, however, believe that the overall architecture is generic and would work well with any light field imaging system once optimized.

Acknowledgements

This work is supported by TUBITAK Grant 114E095.

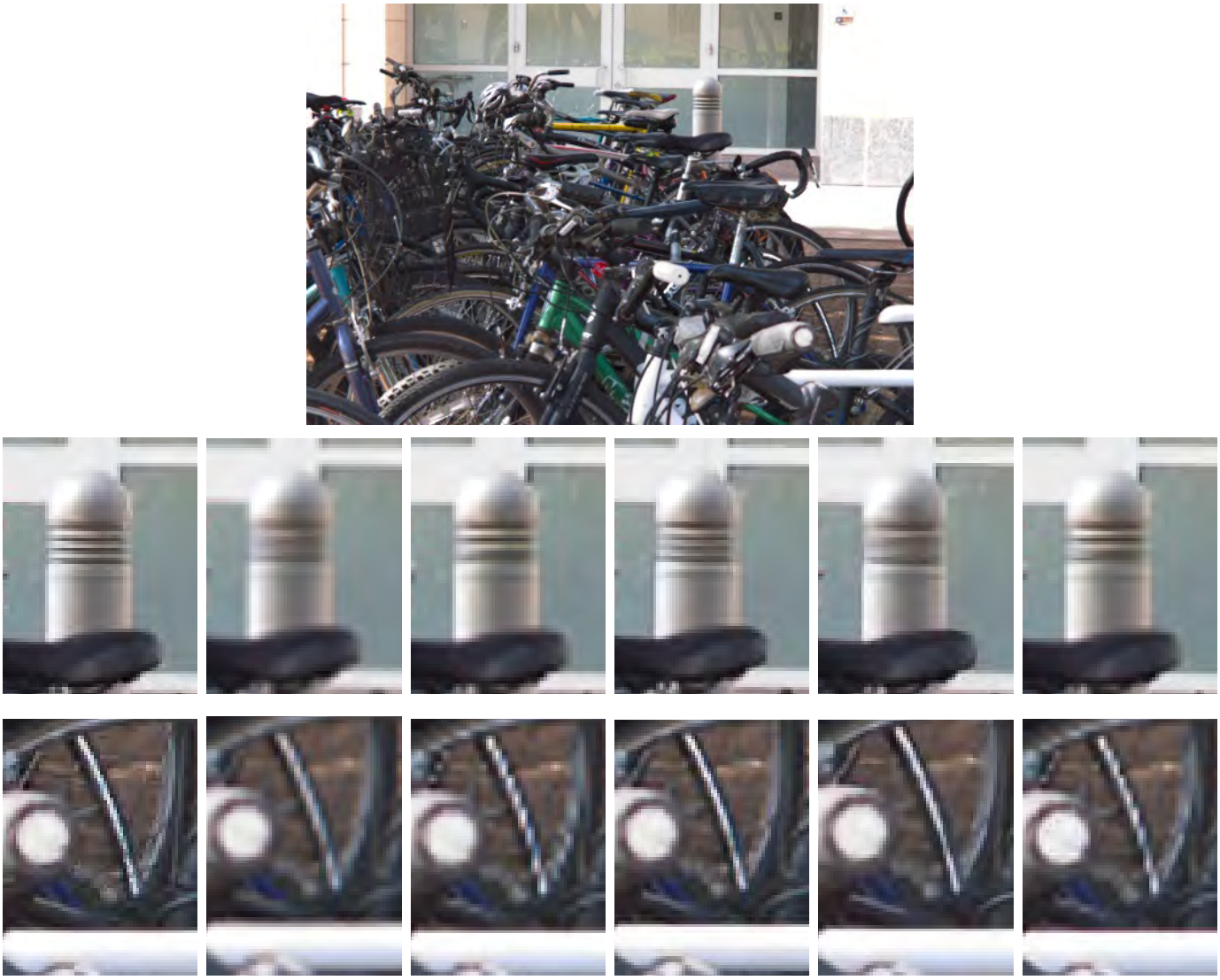
References

[1] G. Lippmann, "Epreuves reversibles. Photographies integrales," *Comptes rendus de l'Academie des Sciences*, pp. 446–451, 1908.

[2] A. Gershun, "The light field," *J. of Mathematics and Physics*, vol. 18, pp. 51–151, 1939.

[3] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," *Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology*, 1991.

[4] E. H. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 99–106, 1992.



(a) Ground truth (b) Bicubic resizing (*imresize*) / 25.99 dB (c) Bicubic interp. / 28.15 dB (d) LFCNN [25] / 24.09 dB (e) DRRN [51] / 33.65 dB (f) LFSR / 29.75 dB

Fig. 12: Visual comparison of different methods. (The worst result image from the dataset is shown here.)

[5] "Lytro, inc." *https://www.lytro.com/.*

[6] "Raytrix, gmbh." *https://www.raytrix.de/.*

[7] M. Levoy and P. Hanrahan, "Light field rendering," in *ACM 23rd Annual Conf. on Computer Graphics and Interactive Techniques*, 1996, pp. 31–42.

[8] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *ACM 23rd Annual Conf. on Computer Graphics and Interactive Techniques*, 1996, pp. 43–54.

[9] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *SIGGRAPH*, 2000, pp. 297–306.

[10] R. Ng, "Fourier slice photography," in *ACM Trans. on Graphics*, vol. 24, no. 3, 2005, pp. 735–744.

[11] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," *ACM Trans. on Graphics*, vol. 25, no. 3, pp. 924–934, 2006.

[12] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *IEEE Int. Conf. on Computational Photography*, 2009, pp. 1–8.

[13] C. Perwass and L. Wietzke, "Single lens 3D-camera with extended depth-of-field," in *SPIE Human Vision and Electronic Imaging*, no. 829108, 2012, pp. 1–15.

[14] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *ACM Trans. on Graphics*, vol. 24, no. 3, 2005, pp. 765–776.

[15] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Trans. on Graphics*, vol. 26, no. 3, pp. Article No. 69:1–12, 2007.

[16] A. Wang, P. R. Gill, and A. Molnar, "An angle-sensitive CMOS imager for single-sensor 3D photography," in *IEEE Int. Solid-State Circuits Conf.*, 2011, pp. 412–414.

[17] V. Boominathan, K. Mitra, and A. Veeraraghavan, "Improving resolution and depth-of-field of light field cameras using a hybrid imaging system," in *IEEE Int. Conf. on Computational Photography*, 2014, pp. 1–10.

[18] X. Wang, L. Li, and G. Hou, "High-resolution light field reconstruction using a hybrid imaging system," *Applied Optics*, vol. 55, no. 10, pp. 2580–2593, 2016.

[19] M. Z. Alam and B. K. Gunturk, "Hybrid stereo imaging including a light field and a regular camera," in *Signal Processing and Communication Application Conf. (SIU)*, 2016, pp. 1293–1296.

[20] T. E. Bishop, S. Zanetti, and P. Favaro, "Light field superresolution," in *IEEE Int. Conf. on Computational Photography*, 2009, pp. 1–9.

[21] S. Wanner and B. Goldluecke, "Spatial and angular variational super-resolution of 4D light fields," in *European Conf. on Computer Vision*, 2012, pp. 608–621.

[22] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, "Modeling the calibration pipeline of the Lytro camera for high quality light-field image reconstruction," in *IEEE Int. Conf. on Computer Vision*, 2013, pp. 3280–3287.

(a) (b) (c) (d)

Fig. 13: Visual comparison of different methods for generating novel views from the HCI dataset. (a) Yoon et al. [52]. (b) Yoon et al. [25]. (c) Proposed LFSR. (d) Ground truth.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] K. Mitra and A. Veeraraghavan, "Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 22–28.

[25] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *IEEE Int. Conf. on Computer Vision Workshops*, 2015, pp. 24–32.

[26] F. Pérez, A. Pérez, M. Rodriguez, and E. Magdaleno, "Fourier slice super-resolution in plenoptic cameras," in *IEEE Int. Conf. on Computational Photography*, 2012, pp. 1–11.

[27] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM Trans. on Graphics*, vol. 34, no. 1, pp. Article No. 12:1–13, 2014.

[28] J. Wu, H. Wang, X. Wang, and Y. Zhang, "A novel light field super-resolution framework based on hybrid imaging system," in *Visual Communications and Image Processing (VCIP)*, 2015, pp. 1–4.

[29] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The J. of Physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[31] C. Dong, Y. Deng, C. Change Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *IEEE Int. Conf. on Computer Vision*, 2015, pp. 576–584.

[32] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 769–777.

[33] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Learning to deblur," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1439–1451, 2016.

[34] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[35] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *IEEE Int. Conf. on Computer Vision*, 2013, pp. 633–640.

[36] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[37] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 769–776.

[38] L. Xu, J. S. Ren, Q. Yan, R. Liao, and J. Jia, "Deep edge-aware filters," in *Int. Conf. on Machine Learning*, 2015, pp. 1669–1678.

[39] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conf. on Computer Vision*, 2016, pp. 649–666.

[40] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomedical Optics Express*, vol. 8, no. 5, pp. 2732–2744, 2017.

[41] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conf. on Computer Vision*, 2014, pp. 184–199.

[42] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *EEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[43] ——, "Deeply-recursive convolutional network for image super-resolution," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645.

[44] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conf. on Computer Vision*, 2016, pp. 694–711.

[45] A. S. Raj, M. Lowney, and R. Shah, "Light-field database creation and depth estimation," *https://lightfields.standford.edu/*.

[46] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Int. Conf. on Multimedia*, 2014, pp. 675–678.

[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Int. Conf. on Artificial Intelligence and Statistics*, vol. 9, 2010, pp. 249–256.

[48] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Annual Workshop on Vision, Modeling and Visualization*, 2013, pp. 225–226.

[49] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. on Graphics*, vol. 35, no. 6, pp. Article No. 193:1–10, 2016.

[50] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, 2014.

[51] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 2790–2798.

[52] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Light-

(a) 15.03 dB (b) 23.54 dB (c) 23.88 dB (d) 20.08 dB (e) 27.80 dB (f) 33.08 dB (g) Ground truth

Fig. 14: Visual comparison of different methods for novel view synthesis. The picture (*"Leaves"*) is taken from Kalantari *et al.* [49]. (a) [50] with disparity [53]. (b) [50] with disparity [54]. (c) [50] with disparity [55]. (d) [50] with disparity [56]. (e) Kalantari *et al.* [49]. (f) Proposed angular SR network. (g) Ground truth.

(a) (b) (c)

Fig. 15: Depth map estimation accuracy. (a) Middle perspective image. (b) Estimated depth map from the input light field with 7x7 angular resolution. (c) Estimated depth map from enhanced light field with 14x14 angular resolution.

(a) (b) (c)

Fig. 16: Depth map estimation accuracy. (a) Middle perspective image. (b) Estimated depth map from the input light field with 7x7 angular resolution. (c) Estimated depth map from enhanced light field with 14x14 angular resolution.

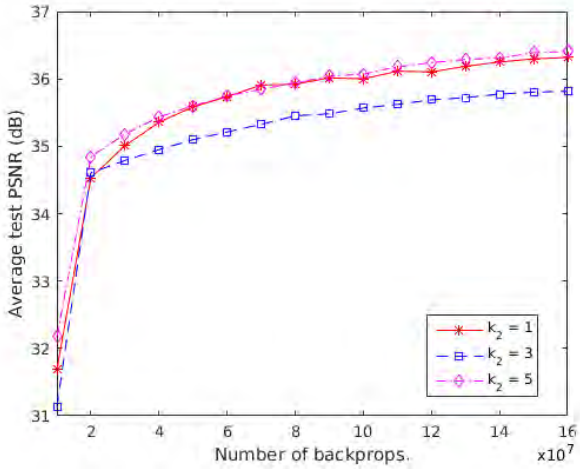


Fig. 17: Effect of the filter size on performance.

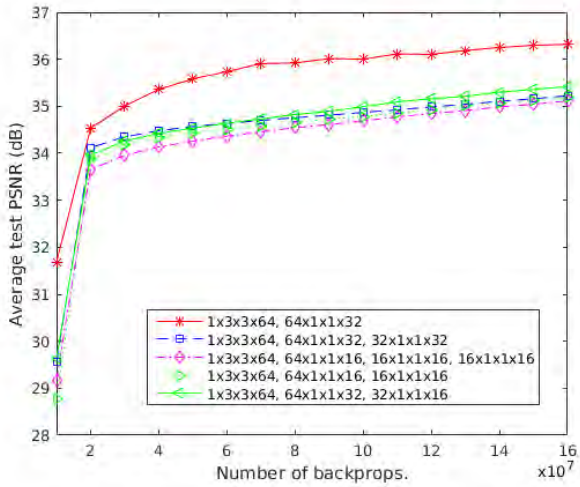


Fig. 18: Effect of number of layers and number of filters on performance.

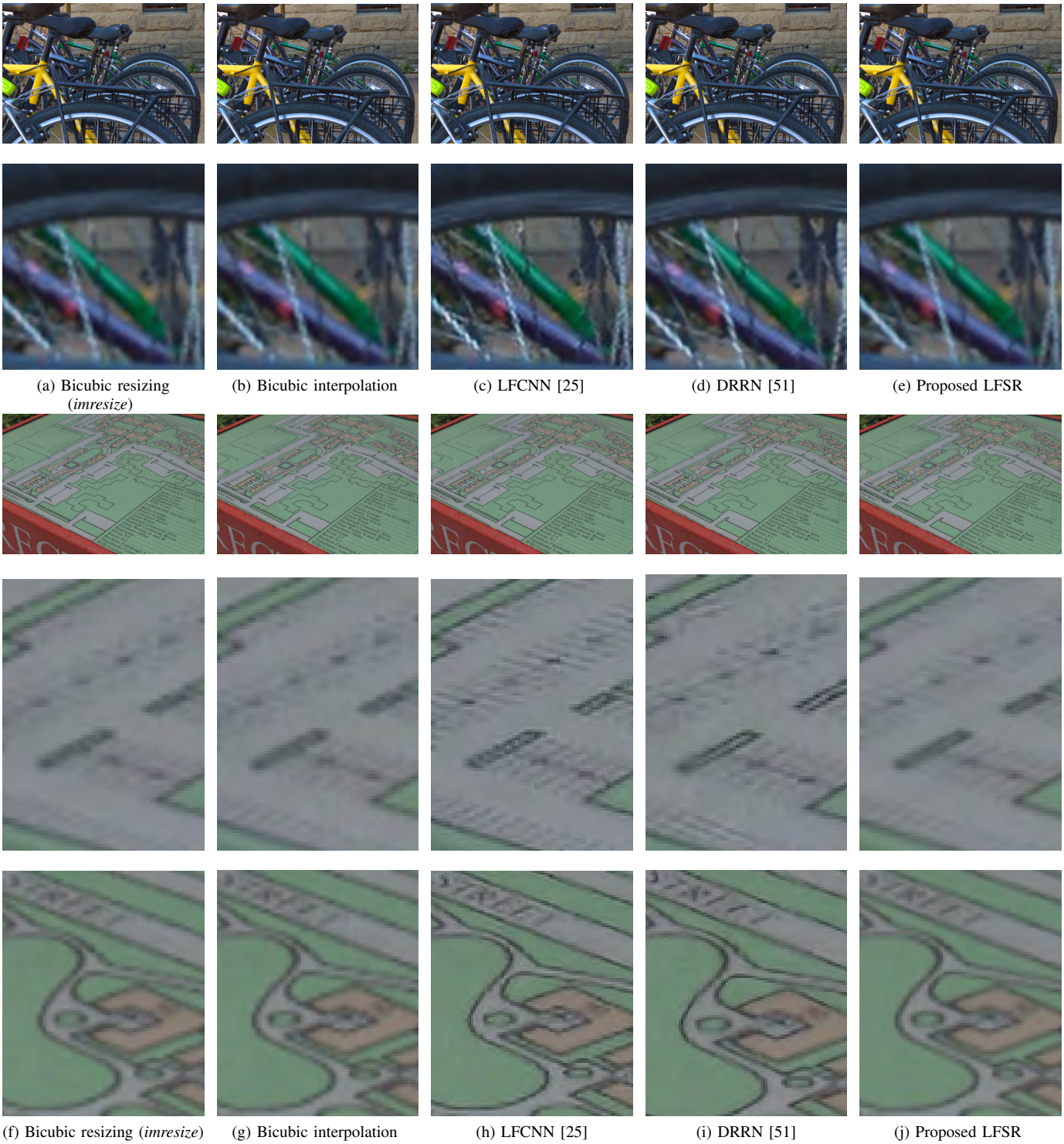
field image super-resolution using convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 848–852, 2017.

[53] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 41–48.

[54] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *IEEE Int. Conf. on Computer Vision*, 2013, pp. 673–680.

[55] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *IEEE Int. Conf. on Computer Vision*, 2015, pp. 3487–3495.

[56] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1547–1555.



(a) Bicubic resizing (*imresize*) (b) Bicubic interpolation (c) LFCNN [25] (d) DRRN [51] (e) Proposed LFSR

(f) Bicubic resizing (*imresize*) (g) Bicubic interpolation (h) LFCNN [25] (i) DRRN [51] (j) Proposed LFSR

Fig. 19: Visual comparison of different methods.